



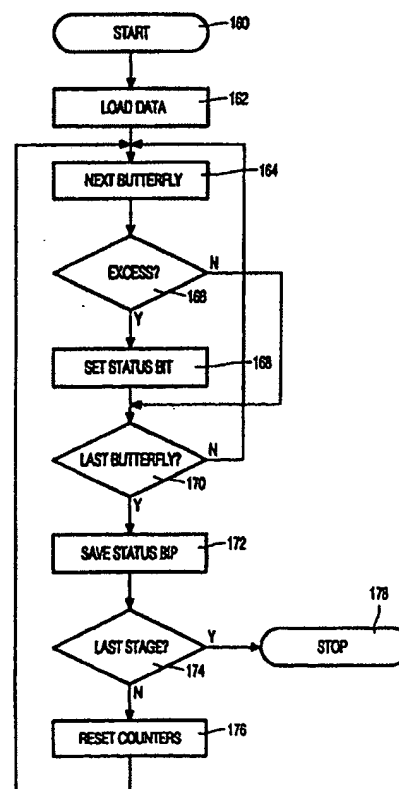
## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>7</sup> : <b>G06F 17/00</b>	<b>A1</b>	(11) International Publication Number: <b>WO 00/07114</b> (43) International Publication Date: 10 February 2000 (10.02.00)
(21) International Application Number: PCT/EP99/04936 (22) International Filing Date: 10 July 1999 (10.07.99) (30) Priority Data: 98202510.8                      27 July 1998 (27.07.98)                      EP (71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL). (72) Inventor: HORSTMAN, Robert, E.; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). (74) Agent: DEGUELLE, Wilhelmus, H., G.; Internationaal Oc- trooibureau B.V., Prof. Holstlaan 6, NL-5656 AA Eind- hoven (NL).	(81) Designated States: CN, JP, KR, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  Published <i>With international search report.</i>	

(54) Title: FFT PROCESSOR AND METHOD WITH OVERFLOW PREVENTION

## (57) Abstract

In a string of modular fixed point data decimated calculation stages, operands of a later stage are formed by results of an immediately earlier stage. Reducing operations are executed to avoid a subsequent overflow condition. In particular, the results are matched to a uniform threshold value, excess of which could lead to said overflow condition in the next decimated calculation stage. If detecting actual excess, the results of the earlier stage are selectively downsized by a uniform factor to avoid a subsequent overflow.



**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

## FFT PROCESSOR AND METHOD WITH OVERFLOW PREVENTION

## BACKGROUND OF THE INVENTION

The invention relates to an FFT processor as recited in the preamble of Claim 1. The invention is also related to a method for performing Fast Fourier Transforms.

FFT is a standard tool used in digital signal processing and has been described  
5 in the book "Discrete-Time Signal Processing" by A.V. Oppenheim and R.W. Schafer, Prentice-Hall 1989, pp. 587-609.

In an FFT processor, an FFT of a block of input samples is calculated by performing subsequently a plurality of basic transform calculations. These basic transform calculations are often called butterfly calculations.

10 During the subsequent calculation of the basic transform calculations, the size of the intermediate results increases continuously. If the intermediate results are represented in a fixed point format it can happen that the intermediate result is larger than the capacity of an actual register in which the intermediate results have to be stored. In this case an overflow occurs. In practice, overflow can occur in every stage, and in the prior art it has been proposed  
15 to normalize the intermediate results in every stage. This can e.g. be done by dividing the intermediate result by a factor of 2 before storing it. This normalization reduces the accuracy of the final results of the complete FFT calculation.

On the other hand, handling the overflow when it actually occurs is quite tedious, because it would necessitate to re-calculate the intermediate result that overflowed, by  
20 execute the basic transform calculation again after normalization of the result of the previous basic transform calculation. This would also require storing various intermediate results longer. Various procedures have been proposed for dealing with the above problem. The publication by Z.A.M. Sharif et al, "Noise analysis for digit slicing FFT", IEE Proceedings-F, No.5, October 1991, pp. 509-512, and a Correspondence thereto by L. Stankovic and R.  
25 Puzovic, Ditto, No. 4, August 1992, p.278, considers the effects of selectively scaling by factors of 1/2 or 1 in each stage, respectively, on the noise caused thereby. The references have not considered a straightforward implementation of the actual scaling itself, but rather deal with software simulations.

## SUMMARY TO THE INVENTION

In consequence, amongst other things, it is an object of the present invention to provide a straightforward fast Fourier processor in which overflow of intermediate results is prevented in a simple way, without having an adverse effect on the accuracy of the complete  
5 FFT calculation.

Therefor a system according to the invention is characterized according to the characterizing part of claim 1.

The present invention is based on the recognition that it is possible to determine from the present intermediate result whether overflow can occur in the next basic transform  
10 calculation to be performed. According to the present invention a scaling by a scaling factor is performed when it is determined that an overflow can occur in the next basic transform calculation. In this way the occurrence of overflow is prevented without performing unnecessary scaling operations.

Preferably, the scaling factor used when it is determined that an overflow can  
15 occur in the next basic transform calculation is a power of two. This makes that the scaling operation can easily be done by performing a shift operation on the intermediate result.

A further embodiment of the present invention is characterized according to the characterizing part of claim 3.

By counting the number of times a scaling occurs, it is possible to obtain a  
20 block floating point representation of the final result of the FFT calculation in a very easy way. In this block floating point representation, the content in the output registers of the calculation means represents the mantissa of the block floating point numbers, and the counted number of scaling operations represented a common exponent (of two) of the floating point representation of the final result of the FFT calculation.

25 A still further embodiment of the present invention is characterized according to the characterizing part of claim 4.

It often occurs in signal processing applications that first an FFT transform is performed on an input signal, that subsequently the result of the FFT transform is subjected to a further processing step, and that the result of the further processing is subjected to an inverse  
30 FFT.

The basic transform calculations in the FFT operation include the multiplication of input samples with a complex constant and the basic transform calculations in the inverse FFT operation include the multiplication of input samples with the conjugate of the complex constant used in the FFT operation.

According to an aspect of the present invention the FFT and inverse FFT operation are combined in one arrangement, in which the needed complex constants are stored only once. The conjugated constants required for the inverse FFT operation are derived from the constants used by the FFT operation by inverting the sign of the imaginary part.

5

## BRIEF DESCRIPTION OF THE DRAWING

These and further aspects and advantages of the invention will be discussed more in detail hereinafter with reference to the disclosure of preferred embodiments, and in particular with reference to the appended Figures that show:

10

Figure 1, a processor arranged for implementing the method;

Figure 2, a flow chart of a procedure according to the invention;

Figure 3, a few vector diagrams used in the butterfly calculation.

## DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

15

Figure 1 shows a processor arranged for implementing the method, by itself being derived from US Patent 4,689,738 to the same assignee and being incorporated by reference herein. Data transport is effected by two parallel buses 20, 22. Items 24, 26, 34, 46, 48, 70, 72, 74, 88, 100, 104, 106, 118, 120, 124, 126, 50, 56 are registers, the latter two via a selection elements. Item 30 is a program counter addressing program memory 28 that can load output register 26, and feed the data buses. Program counter 13 loads from instruction register 26 or from register stack 158. Item 24 is an interrupt address register. Item 90 is a read-only data memory. Items 36, 102 are data memory modules, items 38, 92, 114 address calculation units, items 66, 78 bus selectors. Item 58 is a 16x16 bit multiplier with control register and product register 60. Item 68 is a 40-bit accumulator with add element 64. Item 78 is a bi-directional selector, item 122 an arithmetic and logical unit ALU, item 116 a scratchpad memory, and items 80, 82, 84, 86, 130, 132 are I/O units. For the detailed functionality, reference is had to the cited US Patent. The comparing with the threshold can be implemented at the output of accumulator 68. Another processor type is a so-called Digital Signal Processor (DSP), that features the facility to execute three parallel memory accesses within a single machine cycle. In such high-speed machines, the loosing of a cycle would cause a serious delay.

25

30

Figure 2 is a flow chart of a procedure according to the invention, that will be explained along the calculating of so-called Butterflies in a Fast Fourier Transform FFT. The

FFT comprises a plurality of basic transform operations. In each of the basic transform operations a number of so-called butterfly operations is performed.

There are two alternative algorithms to calculate an FFT. One in the so-called "decimation in time" FFT and the other is the "decimation in frequency" FFT. In the "decimation in time" FFT the butterfly calculation may be represented as rotating a vector in the complex plane, followed by adding another vector thereto. The butterfly operation determines a transformed vector and its conjugate. This can be written as:

$$\begin{aligned} X &= A + B \cdot W_N^k \\ Y &= A - B \cdot W_N^k \end{aligned} \quad (1)$$

The factor  $W_N^k$  is called the twiddle factor, and it represents a complex constant. Its value depends on the number of input samples  $N$ . It is equal to :

$$W_N^k = e^{-j \cdot 2 \cdot \pi \cdot \frac{k}{N}} \quad (2)$$

Alternatively, in a "decimation in frequency" FFT butterflies are calculated which can be represented as adding vectors followed by rotating the result vector in the complex plane. This can be written as:

$$\begin{aligned} X &= A + B \\ Y &= (A - B) \cdot W_N^k \end{aligned} \quad (3)$$

In a fixed point procedure such should not cause an overflow, and as a precaution the calculation of each next basic transform calculation, involving the execution of a number of butterfly calculations, is usually preceded by a division by a factor of 2. A disadvantage thereof is that sometimes this division is unnecessary and will thus cause the unnecessary losing of result bits. A combination of forward and backward transform on an array of 256 real points will in practice cause the losing of about 8 bits. The invention provides an efficient procedure to avoid unnecessary divisions without the necessity to introduce additional cycles.

To avoid overflow, the following condition is necessary and sufficient for "decimation in time" FFT's. A vector with two equal components  $MAX$  can after an arbitrary rotation have no larger component than  $\sqrt{2} \times MAX$ . Adding thereto a second vector with components  $MAX$  will generate a vector with components no larger than a first threshold of  $\sqrt{2} \times MAX + MAX < MAXINTEGER$ , the latter quantity representing the available register length. Therefore, the maximum allowable length of the earlier component is equal to

a threshold of  $\text{MAXINTEGER}/(1 + \text{sqrt}(2))$ . The executing of the earlier calculating stage allows immediate checking this size within its proper cycle and the consequential setting of a signal bit in the status register if the threshold is exceeded. If a basic transform calculation (group of butterflies) is performed, a single such status bit is assigned thereto, that may or may not be set when calculating any single butterfly of the basic transform calculation. After the complete basic transform calculation has been performed, the associated status bit is inspected and if true, the result of the basic transform calculation is divided by a factor of 2 in the next basic transform calculation (group of butterflies) in order to avoid overflow. Counting the number of times the bit has been set for the plurality of basic transform calculations of butterflies will produce a block floating point FFT result in which the result of the counting represents the exponent.

On the other hand, in the "decimation in frequency" FFT, two vectors with components MAX will be added and subsequently rotated in the complex plane. This results in a vector having components which cannot be larger than  $2 \times \text{MAX} \times \text{sqrt}(2)$ .

It has been estimated that about 20-40% of the cycles would be no longer necessary for a 256 point FFT, with respect to another procedure wherein an additional check were executed after each group of butterfly calculations.

Conventionally, in an inverse transform the necessity for two tables is avoided by inverting the sign of the imaginary part of a twiddle factor. The twiddle factor is a number below 1, that indicates the complex vector rotation. Generally, it is represented by two sixteen bit words, of which each first bit indicates the sign and the remainder the size. According to the invention, this may now be executed on a second ALU which procedure will now not influence the above status bit. Another possibility is to introduce temporary saving storage for the above status register. Both features may be straightforwardly introduced into Figure 1.

In particular, the various blocks in Figure 2 represent the following operations. In block 160, the procedure starts by assigning the necessary hardware facilities. In block 162, the input data for the first calculation stage are loaded. Next, in block 164 the first butterfly of the actual basic transform calculation is calculated, and the result is stored. In block 166, the result is compared with the threshold. If the threshold is exceeded, in block 168 the status bit for the actual basic transform calculation is set, if such had not already been effected for an earlier butterfly of the same basic transform calculation. In block 170 it is tested whether the actual butterfly was the last one for the basic transform calculation in question. If negative, the system reverts to block 164 for processing the next butterfly. If positive, in block 172 the status bit is saved, for in the next multiplying stage controlling a division by 2. For this

purpose, the output of the multiplier is followed by a shift mechanism not shown separately in the Figure.

In block 174 it is tested whether the actual basic transform calculation was the last one to be performed. If negative, the system resets the necessary butterfly counters, etcetera, and reverts to block 164 for executing the butterflies of the next basic transform calculation. If during the execution of the most recent basic transform calculation the status bit was set, The output samples of said basic transform calculation is normalized by dividing by the decimating factor of two before the next basic transform calculation is executed. This is implemented by left-shifting over one binary position at the multiplying in question. In block 10 178, the system terminates operation, by outputting all results, inclusive of the number of status bits that had been set, for so allowing a floating point to be used subsequently. For example, if a 256 point transform had five bits set, the floating point coefficient is incremented by 5. In this way, a fixed point machine provides the correct block floating point output data.

Figures 3A, 3B show a few vector diagrams used in the butterfly calculation. In 15 Figure 3A, for a "decimating in time" FFT, a first vector has two components with values **MAX** each. No larger components can occur. The size of this vector can thus be no larger than **MAX sqrt(2)**. This vector is added in the calculating of the butterfly to another vector component that can have no greater value than **MAX** again. All other sizes and vector directions lead to lower values of the result. Therefore, the overall result can be no larger than 20 **MAX x (1 + sqrt(2))**. Therefore, the test threshold for the result is the inverted value of this expression.

In Figure 3B, for a "decimating in frequency" FFT, two first vectors have components with values **MAX** each; no larger components can occur, and the case shown that they have parallel directions represents a worst case. The size of the combined vector can thus 25 be no larger than **2 x MAX**. This vector is rotated in the calculating of the butterfly so that the largest component of the rotated result can have no greater value than **2 x MAX sqrt(2)**. All other sizes and vector directions lead to lower values of the result. Therefore, the test threshold for the result is the inverted value of this expression.

## CLAIMS:

1. Fast Fourier processor comprising calculation means for performing a plurality of subsequent basic transform calculations, said calculation means comprise scaling means to prevent overflow of a result of said basic transform calculations, characterized in that the fast Fourier processor comprises means for determining whether a result of at least one of the basic transform calculations exceeds a threshold value, and in that the scaling means are arranged for scaling said result of said at least one of the basic calculations by a scaling factor smaller than one if said result exceeds the threshold value.
- 5 2. Fast Fourier processor according to claim 1, characterized in that said scaling factor equals two.
3. Fast Fourier processor according to claim 1 or 2, characterized in that the fast Fourier processor comprises counting means for counting the number of times a scaling takes place during performing said plurality of basic transform calculations, and in that said number of times represents an exponent of a block floating point representation of a final result of the plurality of subsequent basic calculations.
- 15 4. Fast Fourier processor according to one of the previous claims, characterized in that the fast Fourier processor is arranged for calculating a first transform and its inverse transform, in that for the first transform the basic transform calculations comprise a multiplication of an input signal with a first complex constant, in that for the inverse transform the basic transform calculations comprise a multiplication of an input signal with a second complex constant which is the conjugate of the first constant, in that the fast Fourier processor comprises inverting means for calculating the second constant from the first constant by
- 20 inverting the sign of the first constant.
5. Fast Fourier processor according to claim 4, characterized in that the inverting means comprise different means than the calculation means.

6. Fast Fourier processor according to claim 4, characterized in that the inverting means comprise a part of the calculation means, and in that the Fast Fourier processor comprises storage means for temporarily storing an indicator indicating whether a result of a most recently executed basic transform calculation exceeded said threshold value.

5

7. Fast Fourier transforming method comprising performing a plurality of subsequent basic transform calculations and preventing overflow of a result of said basic transform calculations, characterized in that the method comprises determining whether a result of at least one of the basic transform calculations exceeds a threshold value smaller than the maximum value of said result, and in that the method comprises scaling said result of said at least one of the basic calculations by a scaling factor smaller than one if said result exceeds the threshold value.

10

8. Method according to claim 7, characterized in that said scaling factor equals two.

15

9. Method according to claim 7 or 8, characterized in that the method comprises counting the number of times a scaling takes place during performing said plurality of basic transform calculations, and in that said number of times represents an exponent of a block floating point representation of a final result of the plurality of subsequent basic calculations.

20

10. Method according to one of the claims 7, 8 or 9, characterized in that the method comprises calculating a first transform and its inverse transform, in that for the first transform the basic transform calculations comprise a multiplication of an input signal with a first complex constant, in that for the inverse transform the basic transform calculations comprise a multiplication of an input signal with a second complex constant which is the conjugate of the first constant, in that the method comprises calculating the second constant from the first constant by inverting the sign of the imaginary part of the first constant.

25

1/3

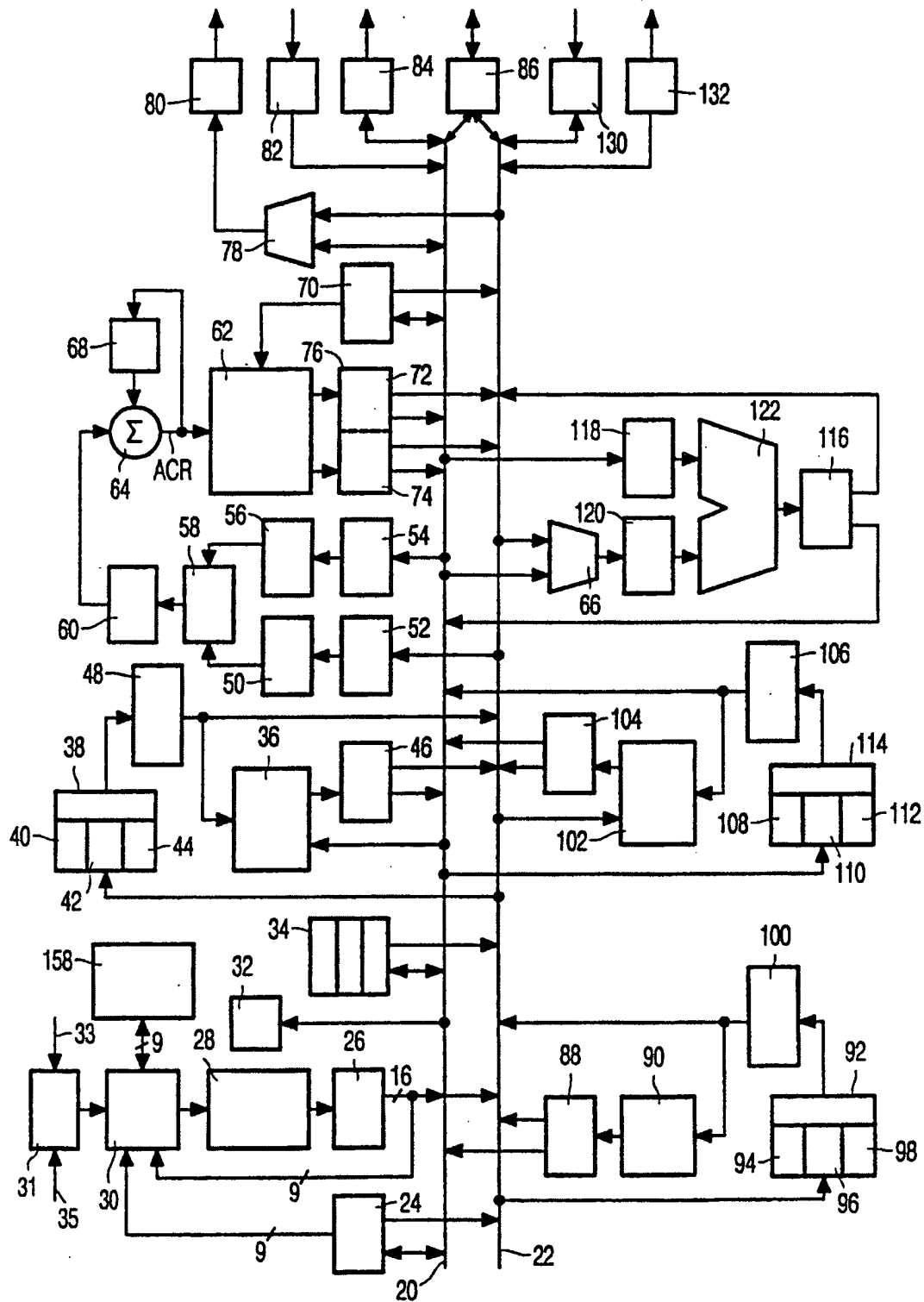


FIG. 1

2/3

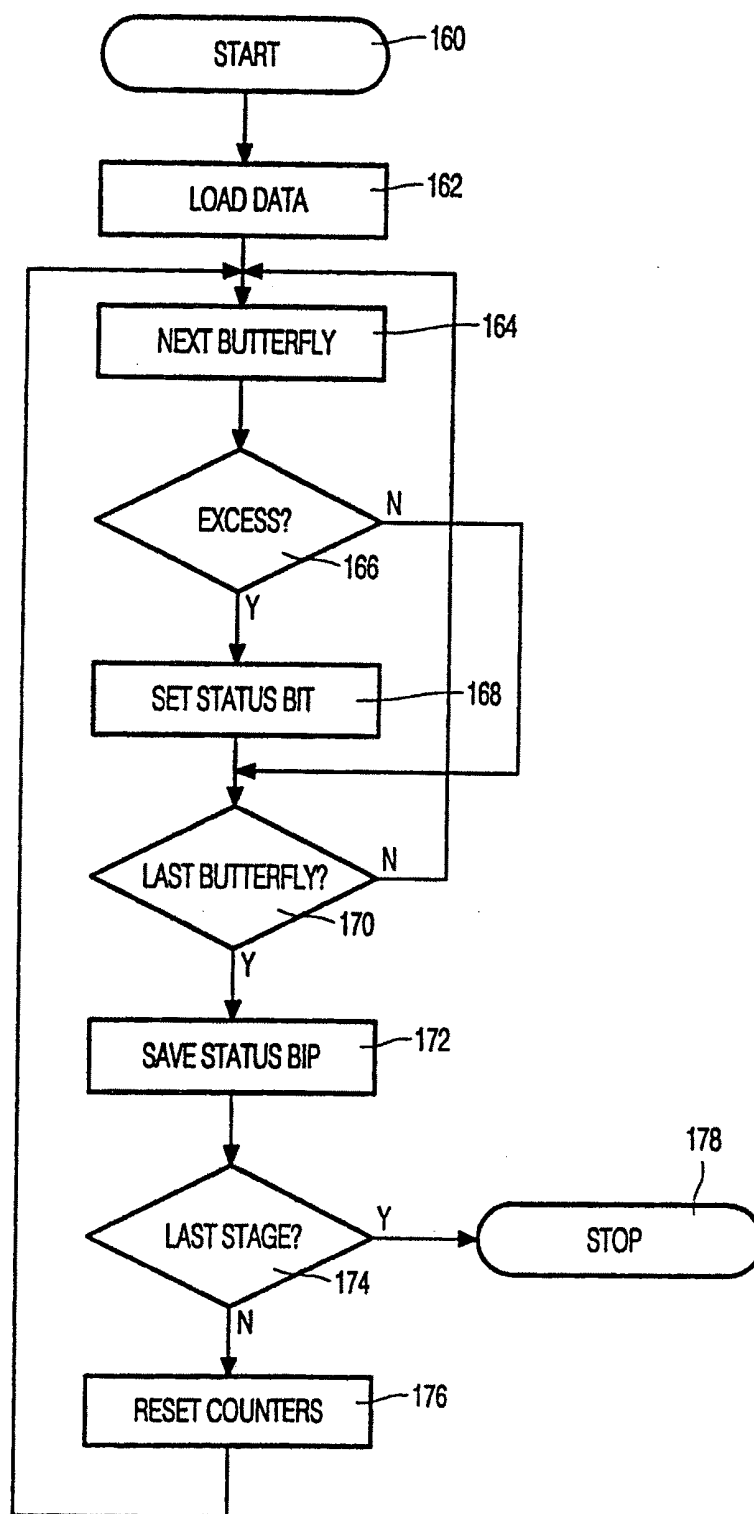


FIG. 2

3/3

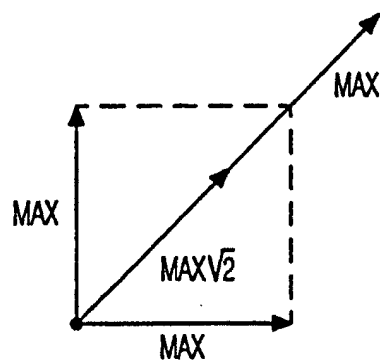


FIG. 3A

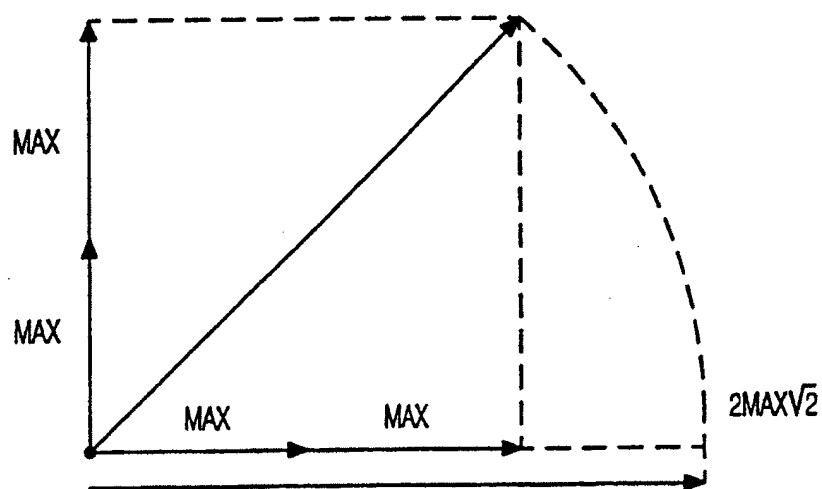


FIG. 3B

# INTERNATIONAL SEARCH REPORT

Internat. Application No

PCT/EP 99/04936

## A. CLASSIFICATION OF SUBJECT MATTER

IPC 7 G06F17/00

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 4 501 149 A (KONNO JUNICHI ET AL) 26 February 1985 (1985-02-26) column 6, line 42 - line 64 claims 3,4	1-10
A	US 4 872 132 A (RETTETTER REFAEL) 3 October 1989 (1989-10-03) claims 1-6 column 2, line 5 - line 50	1-10
A	US 5 481 488 A (XU JIASHENG ET AL) 2 January 1996 (1996-01-02) abstract; claim 1	1-10
A	EP 0 155 660 A (HEWLETT PACKARD CO) 25 September 1985 (1985-09-25) abstract; claim 1	1-10

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubt on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

11 October 1999

Date of mailing of the international search report

19/10/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Filloy García, E

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/EP 99/04936

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 4501149 A	26-02-1985	JP 1021461 B	21-04-1989
		JP 1538893 C	16-01-1990
		JP 59079852 A	09-05-1984
		DE 3339288 A	03-05-1984
		DK 494483 A, B,	30-04-1984
		GB 2129560 A, B	16-05-1984
US 4872132 A	03-10-1989	NONE	
US 5481488 A	02-01-1996	NONE	
EP 0155660 A	25-09-1985	DE 3585449 A	09-04-1992
		JP 1828149 C	28-02-1994
		JP 5030327 B	07-05-1993
		JP 60210016 A	22-10-1985
		US 4750145 A	07-06-1988

THIS PAGE BLANK (USPTO)